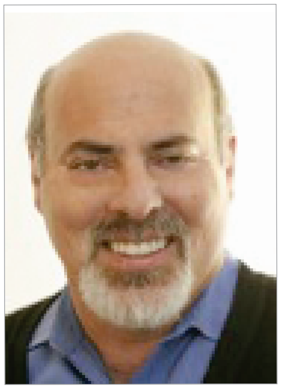
Analytics in Database:

Simplify the Effort, Power Up the Capability

Neil Raden
Hired Brains Research
February 2010

ABOUT THE AUTHOR



Neil Raden, based in Santa Barbara, CA, is an active consultant and widely published author and speaker and also the founder of Hired Brains, Inc., http://www.hiredbrains.com. Hired Brains provides consulting, systems integration and implementation services in Business Intelligence, Decision Automation and Business Process Intelligence for clients worldwide. Hired Brains Research provides consulting, market research, product marketing and advisory services to the software industry.

Neil was a contributing author to one of the first (1995) books on designing data warehouses and he is more recently the co-author with James Taylor of Smart (Enough) Systems: How to Deliver Competitive Advantage by Automating Hidden Decisions, Prentice-Hall, 2007. He welcomes your comments at nraden@hiredbrains.com or at his blog at Intelligent Enterprise magazine at http://www.intelligententerprise.com/blog/nraden.html.

EXECUTIVE SUMMARY

Beyond reporting and manipulation of aggregated data through reports or dashboards, the need for more advanced analytics is being felt. Previously the reserve of highly specialized "quants," analytics is moving to center stage as the next "must have" technology. Unfortunately, current tools are not optimized for either a greater level of activity or for the massive data sets now routinely created in data warehouses.

While the cost of data warehouse appliances has dropped to under $20k/terabyte, the cost of running an advanced analytics department has increased. To put it in perspective, it is about 2000 times LESS expensive to build a data warehouse today (per byte) than it was just 10 years ago, but of course, the data volumes have also increased. Nevertheless, the cost of building a medium-to-large data warehouse is still lower, even in absolute terms. As a result, more data is gathered and the demand to analyze it, to use it to advantage, has increased sharply.

In order to leverage this rich trove of data, better solutions are demanded. At the very least, moving quantitative methods into the data warehouse make sense if only because they eliminate the expensive, time consuming, latency-causing extraction and movement of data from a data warehouse to an analytics platform. Unfortunately, the database software market has not responded. Even the so-called analytical database market has been slow to support this idea. But these capabilities are disruptive and Netezza is pairing its database appliance technology with not only statistical routines, but an entire application development environment for advanced statistical applications utilizing the same in-place Netezza platform. Combined with the simplicity, speed, scalability, cost position and reliability, Netezza is providing the needed platform to effectively deploy advanced analytics throughout the organization.

WHERE THE DATA IS

The infamous bank robber Willie Sutton (1901-1980), when asked why he chose bank robbing as a profession, had a simple answer: "That's where the money is." Sigmund Freud might have written a book about Sutton's motivation, but it's clear to see there is real merit in simplicity. A few years ago, Netezza saw the advantage of simplicity by applying some Willie Sutton thinking ("that's where the data is") and bundling a massively scalable computing platform, a relational database tuned to both the physical platform and the task of analytics and integrated storage devices to handle multiple terabytes of data, all in one package. By creating the first data warehouse appliance, an entirely new market in the analytic appliance field, Netezza broke the mold of databases, servers and storage sold and integrated as separate components. At the time, relational databases were standalone software that relied on third-party servers and external storage devices.

It made perfect sense to build large data warehouses housed in a single appliance. It reduced the load on the least scalable part of the architecture, the network. It placed the logic satisfying queries as well as loading and accessing data inside the appliance next to the physical storage devices. It used some creative hardware and software to allow for extreme speed, scale and reliability in operations to physical storage. The goodness of fit of this solution is evidenced by both the success of Netezza in the market as well as the number of copycats that have emerged. But the work wasn't finished.

Data warehousing and business intelligence as disciplines are in their third decade and as a result, they bear resemblance to their original characteristics in name only. Just over ten years ago, a data warehouse could sit on a server with 1/10,000th the processing power, and able to store (at great expense) perhaps 20-50 gigabytes of information. The cost of this environment, priced separately for database software and physical storage could easily exceed $1,000,000. Performance was a constant problem, requiring the full-time attention of one or more DBA's. Because of all these limitations, the demand for services from the data warehouse was dampened and limited to reporting and a small percentage of analysis, using OLAP tools for example.

But there was always a small group of people in organizations whose needs went beyond analysis to advanced analytics. Given the limited capabilities and resources of the data warehouse platforms, the standard practice was for analysts to extract data from the warehouse in bulk, manipulate in another environment until it was readable by a statistical and/or data mining tool, then run the procedures remotely. The drawbacks were (and still are in many cases) extreme. The network was forced to handle this traffic in data that had already been moved from other places into the warehouse. Often, analysts were forced to use only a sample of the data because of constraints of data warehouse and network loads. Most importantly of all, it consumed far too much valuable time of the analysts who had more important things to do than moving, preparing and storing data.

[1] The quote is so famous that there is a "Willie Sutton Rule" in the field of Activity Based Costing (ABC) in Management Accounting that ABC should be applied "where the money is," the source of the highest costs.

So the logical, Willie Sutton solution would be to move the analytical processing to the data warehouse platform. As the capabilities of database engines for large scale analytics increased, the demand for these services increased. With the release of his article, "Competing on Analytics," (Harvard Business Review, January, 2006) and subsequent book of the same title, Tom Davenport, a popular business writer, more or less raised everyone's consciousness about the need for and value of advanced analytics in organizations.

ANALYTICS

There is no single accepted definition for the "term" analytics, often modified with the adjective "advanced." In general, anyone reviewing data to determine either what happened or what is likely to happen could be said to be doing analytics. Some argue that OLAP (Online Analytical Processing), and interactive navigation though data arranged in hierarchies allowing for "drill down" or "pivot" operations is a form of analytics. But popular authors like Tom Davenport I define analytics as the application of quantitative methods, to create predictive models, provide optimization services and to expose senior management to the effects of these efforts.

The terms "analytics" and "advanced analytics" are often used interchangeably as they lack precision in definition, but in general, creating an analytical environment includes the following activities:

- **Data cleansing:** When it comes to cleanliness, clean is in the eye of the beholder. What is considered clean in a data warehouse may not be suitable for certain analytical efforts. Missing data, outliers, miscoded attributes – all need to be dealt with. In current practice, this is often a necessary but time consuming manual activity performed by the analyst that doesn't add significant value.

- **Data exploration:** Data exploration is the process of examining a set of data to gain understanding. This often starts with the extracting of values such as counts, distinct counts, mean, median and variance. A fairly simple technique, classification, clumps the data into predefined buckets, such as 10-year age brackets or income by strata. A more mysterious process, often called cluster analysis, clumps the data by discovering unseen or undetected relationships in the dataset and performing its own classification.

- **Visualization:** When analyzing billions of points of data (or trillions as is rapidly becoming the norm), aggregated statistics (mean, dispersion, correlation, etc.) are very useful, but nothing really compares to a graphical visualization of the data, especially if the visualization is interactive, meaning, the analyst can change the elements or orientation of the visualization in real-time.

[2] Keep in mind that analysts include many people not working in commercial organizations in fields such as engineering, astronomy, biology, remote sensing or economics.

- **Attribute extraction/problem representation:** Data for analytics is, or at least should be, richly attributed. Consider a credit file: there may be hundreds of pieces of information about an individual in addition to payment history. Because the computational cost of sifting through all of these attributes is excessive, it makes sense, for a particular analysis, to remove from consideration those attributes that are known to carry little weight or are so closely related to other attributes that their contribution is not needed. Similar to attribute extraction, dimensionality reduction uses Principal Components Analysis to create a linear mapping of the values to a lower level of dimensionality, creating a correlation matrix of the dimensions, computing eigenvectors of this matrix and determining essentially, what can be left out with minimal data loss.

- **Predictive analytics:** Predictive analytics makes predictions about the future. The future may be weeks or months or years away, or may be only a millisecond away. What qualifies analysis as predictive is that it synthesizes likely outcomes that have yet to occur. This is more or less the opposite of descriptive analytics that describe what has already occurred. An example is demand forecasting, a critical piece in the supply chain. Because of the uncertainty of demand across the time horizon of manufacture, order, ship and store cycles, especially when out of stock situations are unacceptable, such as medical supplies, defensive stock positions are taken that cause excess inventory costs. By being able to more accurately predict actual demand, inventories can be reduced at both the customer and supplier sides.

- **Scoring:** Scoring is a form of predictive analytics in that it assigns a value to each record in order for other models to take actions. The FICO credit score is a good example of this. FICO is a very complex model, whose inner workings are a trade secret, but if you think about it, emergency room triage is a much simpler (and desperate) form of scoring: those to save, those that can't be saved, those that can wait.

- **Stochastic process:** When it comes to predictive processes, deterministic models calculate outcomes based on a predetermined model. Stochastic processes are random, where the input variables are determined at run time using a probability distribution and random number seed. They are typically repeated many thousands or even millions of times. Time series analysis and Monte Carlo simulations are good examples of this and used widely in predicting complex events whose interactions may not fully understood.

- **Reporting and publishing:** Though not often mentioned as a part of advanced analytics, the outcome of analytic analysis has to be presented and explained to others. Reporting requirements for analytics are somewhat different than the row-and-column of financial, sales and marketing reports. They may include more graphics and visualization, the use of special characters and the arrangement of data in other than symmetric formats. Creating these reports can be taxing on the analyst, the system or both.

- **Maintaining models, QA, re-builds and revisions:** Descriptive and predictive models, once vetted, need attention on an ongoing basis as all of the elements that affect them are subject to constant change. Good analytical tools provide much more than quantitative algorithms. The analytical process requires something like an integrated development environment (IDE) for development, versioning, maintenance, enhancement, governance and monitoring of dependencies.

DRAWBACKS

Analytic technology exists on two levels. On the first, it is highly theoretical and practiced only by those with an academic and/or research orientation. Time is not typically of the essence. On the other hand, analytics applied in commercial environments are under pressure to both perform in a timely manner and be able to explain their results to those who do not have the same level of skill and comfort with quantitative methods. In addition, the industry problems are considerably more varied. Combining this with the explosion of data available to analytics professionals in a commercial (or government) enterprise, existing solutions, even data warehouse appliances like Netezza, are not sufficient. They may be able to store and process terabytes of data easily and quickly, but they cannot perform the analytical processes needed. Instead, data is extracted from them and placed on a server with the appropriate analytical engine. In practice, these servers are considerably less powerful than a data warehouse appliance.

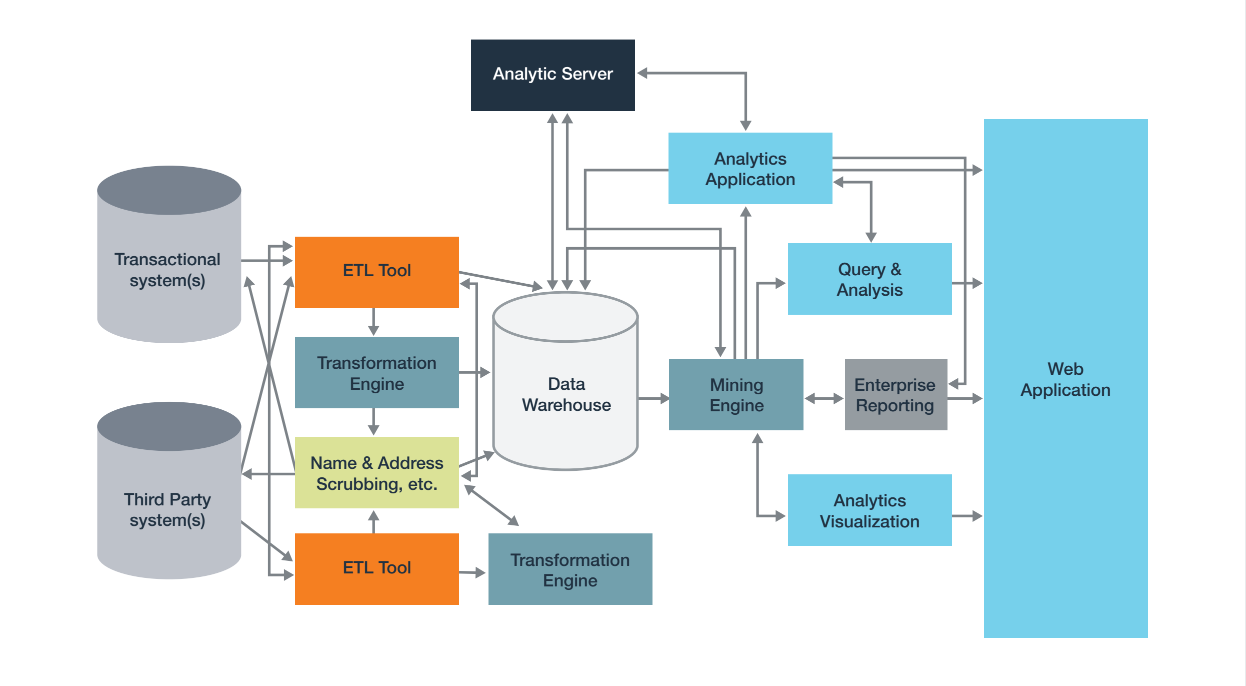
IN-DATABASE VS. TRADITIONAL STAT METHODS:

Yesterday:

1. Buy million$ in statistical software licenses
2. Spend days iterating the correct hundred GB data transfer from the DB to your stat server
3. Run analytics (rinse and repeat until correct)
4. Report done 8 days after "go" date

Tomorrow:

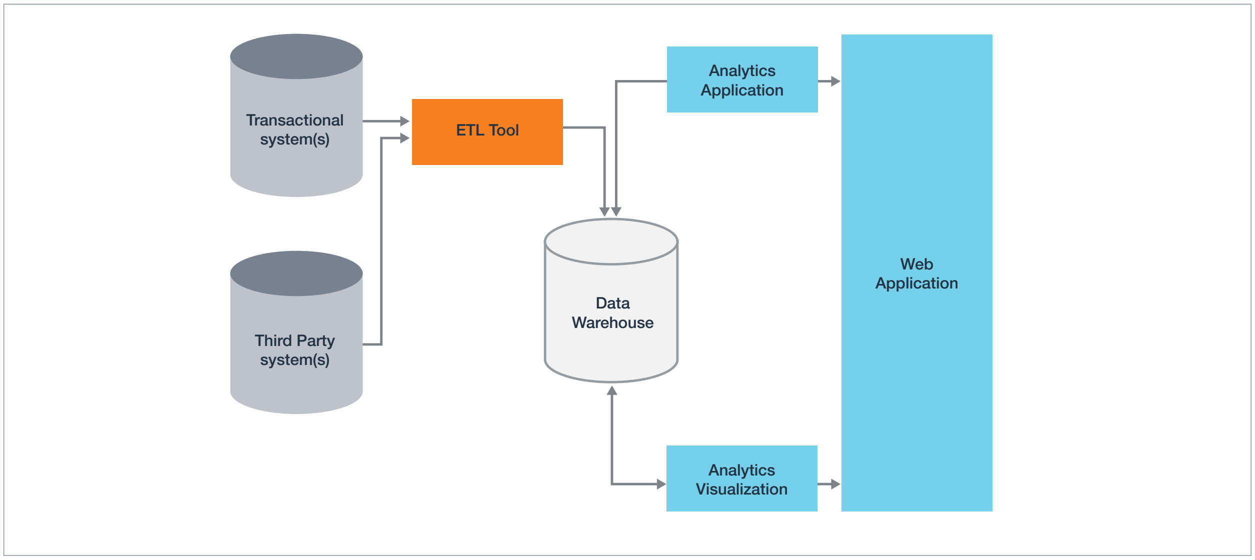
1. Iterate pull and analyses locally on Netezza server
2. Report done same day as "go" date



Another drawback is that the statistical packages themselves are not a complete solution. Often, applications are written around them, in Java, PHP, Python, etc., causing yet another shift of data and processing and a de-coupling of the statistical and application aspects, leading to increased maintenance potential for error and latency.

NEEDED IMPROVEMENTS

The solution is to create a true analytics platform that is a monster data manager coupled with all of the analytical supercomputing processing capabilities that is fully integrated with a rich suite of analytics and development tools in a single environment that minimizes data shifting and provides equivalent processing power and availability to the entire effort of data integration to analytics to presentation.



CAPABILITIES OF A TRUE ANALYTICS PLATFORM

Using the Willie Sutton method, it makes perfect sense to locate the processing where the data is. And when the data is voluminous and complex, it is the only alternative. A powerful advanced analytics platform that supports all the advanced analytics activities must start with a powerful data warehouse appliance, but has to accommodate the downstream analytical work to avoid needless movement of data and a more coherent platform to support the process end-to-end. Moving the processing next to the data is a good first step, such as embedding quantitative functions or even existing statistical packages into the data warehouse appliance, but it is only a partial solution. Other capabilities needed are:

- **Tight Integration of the analytics and analytic applications that exploits the parallel architecture:** Creating parallel applications for analytics and models is a difficult task but the performance and scalability benefits enable near real-time analytic processing. To be effective, the platform has to ease the implementation of analytic applications into the parallel architecture.

- **Flexible and rich analytics development environment**: While analytics have been around for over 30 years there is no standard language, environment or framework for analytics in existence today. Therefore an analytics platform must be flexible and support the needs of a diverse set of analytics users. While the emerging defacto analytic language appears to be the open source R there is a deep history of analytics developed in everything from Fortran to C to Java and many others. Similarly open source and vendor supplied IDEs and ADEs (analytic development environments) have been used over the years. A viable analytics platform must support and allow additional languages and tools to easily integrate into their platform.

- **Open source analytics support:** Beyond just the cost advantages of open source, there are large, innovative and productive communities contributing to the portfolio of analytical procedures. If they cannot be easily plugged into the analytic appliance, the cost and innovation advantages are lost.

DIFFERENTIATORS FOR A TOTAL ANALYTIC PLATFORM

A total solution for analytics will provide the greatest level of effectiveness if it combines the elements of simplicity, scalability and performance. After many years of practice, there remains a notion that data belongs in a database and other analytical processes belong elsewhere. The more platforms and processes involved, the greater the complexity of establishing, maintaining and enhancing an analytical solution. Though analytical databases of today are capable of handling vast amounts of data, the time to extract data to other platforms and the limited bandwidth of network resources adds a measurable amount of latency to the process. Finally, servers for analytical software are rarely scaled at the level of a database appliance and cannot offer the processing speed and throughput of an integrated analytic platform.

SIMPLICITY

Current architectures for analytics are cumbersome, redundant, expensive and complex. A quantitative analyst has to master, at the very least, SQL, a statistical package such as SAS or SPSS, presentation tools either in classic BI packages, Excel or third-party tools, or all of them. Because of this, analysts spend a great deal of their time doing basic data work instead of devising and analyzing models, the role they are trained and compensated for. But the appliance approach alleviates most of this effort by not only housing the data and the logic to process it, but by providing state-of-the-art performance, reliability, load balancing and availability. The appliance approach also reduces cycle time from model development to model deployment, improving the productivity of the modelers. It can reduce the complexity of the jobs themselves by, for example, embedding the statistical or data mining routines into SQL queries thereby reducing two (or more) steps to just one.

Another useful consequence is the ability to actually process all of the data, not just a sample. With separate statistical servers, it isn't feasible to move a terabyte or more so analysts typically spend time creating (hopefully) unbiased samples. With integrated analytics, and the ability to scan huge amounts of data quickly, it is possible to run models against the entirety of the data, not just a sample.

SCALABILITY

The most obvious benefit of combined database/analytics on a massively parallel platform is the ability to run more models in parallel model execution. The machine time and cycle time in current environments with the engines separated consumes too much of the available clock. Scalability, especially the ability to achieve linear scalability by adding blades, enables the analytic workload to increase dramatically.

Because analysts will spend less time tending to data issues, they will be able to do more modeling. This accentuates the highest value added by an analyst. In addition, because larger data volumes can be handled, more insight into historical data and transactional details is possible. This of course leads to better insights by analyzing deeper levels of transactions or longer history.

In addition to being able to analyze more data faster, the analytic appliance holds the promise of being able to amplify an organization's innovation by posing and answering more complex business problems, and creating comprehensive solutions that are not possible on other platforms. The consistency of the architecture and the lack of hand-off's, extracts and interfaces between systems promotes more comprehensive models and more of them.

PERFORMANCE

Performance has to be measured as analytic throughput, not just the speed of a machine. To get that throughput, the best strategy is to reduce redundant steps and latency. Once that happens, it frees the analyst for other opportunities, for example, faster results with greater data volumes, enabling a deeper understanding of the causes and effects reflected in the data.

Overall, doing more from a single platform is more logical than ever if the platforms are scalable, cost-effective and well-designed.

CONCLUSION

The Netezza TwinFin i-Class™ is not just an evolution of the Netezza product line, it represents a new market segment, the analytics platform. It will disrupt traditional database vendors who have still not gotten their general purpose relational databases to perform very well, and it will disrupt the nascent data warehouse appliance market that it created. Standalone high-performance data warehouse appliances still have a valuable role to play, but where advanced analytics are needed, they are inadequate. And merely tacking on a statistical package is likewise inadequate. Serious analytics is as much a software engineering discipline as developing operational systems. It has to be supported by a complete, balanced IDE in order for analysts not to be swamped by the capabilities. Building 10 times as many models and running them 20 times more often requires support tools to manage the process, and results from model runs will ultimately be piped to other applications. The Netezza TwinFin i-Class is designed to step up to those requirements.

Hired Brains, Inc.
1415 Kenwood Road
Santa Barbara, CA, 93109
Phone: 805-962-7391